Between-speaker variability in temporal organizations of intensity contours

He, Lei ; Dellwo, Volker

Abstract: Intensity contours of speech signals were sub-divided into positive and negative dynamics. Positive dynamics were defined as the speed of increases in intensity from amplitude troughs to subsequent peaks, and negative dynamics as the speed of decreases in intensity from peaks to troughs. Mean, standard deviation, and sequential variability were measured for both dynamics in each sentence. Analyses showed that measures of both dynamics were separately classified and between- speaker variability was largely explained by measures of negative dynamics. This suggests that parts of the signal where intensity decreases from syllable peaks are more speaker-specific. Idiosyncratic articulation may explain such results.

DOI: https://doi.org/10.1121/1.4983398

Between-speaker variability in temporal organizations of intensity contours

Lei He and Volker Dellwo

Citation: The Journal of the Acoustical Society of America **141**, EL488 (2017); doi: 10.1121/1.4983398
View online: http://dx.doi.org/10.1121/1.4983398
View Table of Contents: http://asa.scitation.org/toc/jas/141/5
Published by the Acoustical Society of America

Between-speaker variability in temporal organizations of intensity contours

Lei He[a] **and Volker Dellwo**
Phonetics Laboratory, Institute of Computational Linguistics, University of Zurich, Andreasstrasse 15, CH-8050 Zurich, Switzerland
lei.he@uzh.ch, volker.dellwo@uzh.ch

Abstract: Intensity contours of speech signals were sub-divided into positive and negative dynamics. Positive dynamics were defined as the speed of increases in intensity from amplitude troughs to subsequent peaks, and negative dynamics as the speed of decreases in intensity from peaks to troughs. Mean, standard deviation, and sequential variability were measured for both dynamics in each sentence. Analyses showed that measures of both dynamics were separately classified and between-speaker variability was largely explained by measures of negative dynamics. This suggests that parts of the signal where intensity decreases from syllable peaks are more speaker-specific. Idiosyncratic articulation may explain such results.

[AL]
Date Received: December 12, 2016 **Date Accepted:** April 28, 2017

1. Introduction

Source signals, vocal tract resonances, and articulatory movements are the essential processes of speech production. Each of these processes encodes speaker-specific information.[1] This study investigated how between-speaker differences are reflected in temporal organizations of intensity contours in terms of intensity dynamics. Intensity dynamics were defined as the speed of increase in intensity from an amplitude envelope trough point to a consecutive peak point (henceforth, positive dynamics) and the speed of decrease in intensity from a peak to a consecutive trough point (henceforth, negative dynamics).

Speaker idiosyncratic characteristics in both glottal vibrations and vocal tract resonances have been extensively studied in forensic phonetics and automatic speaker recognition (see Eriksson[3] and Kinnunen and Li[4] for reviews). Far less attention has been paid to the temporal characteristics of speech that are a result of the movements of the articulators over time.[5–7] The rationale of these studies is that articulatory movements are comparable to other domains of human movements (e.g., gait and typing) where individual differences are conspicuous.[3,5–7] Such individualities are related to both individual neurological dispositions, which constrain the motor control over the respective body parts,[8] and ontogenetic anatomical characteristics of moving body parts, which shape their biomechanical properties.[9] As a specialized domain of human motor behavior, articulation also reflects speaker individualities because of anatomical idiosyncrasies of the articulators and the way speakers acquired control over them. These result in speaker-specific articulatory kinematics, including velocity, acceleration and spatial displacement.[10,11] Such kinematic characteristics are assumed to be the reason for speaker-specific production of prosodic duration[5,6] and intensity variabilities.[7,12] The present research underlies the assumption that the intensity contour shape might be closely related to the articulatory movements responsible for the changes of mouth opening area in an utterance. Such a view is supported by Summerfield[13] who held that the amplitude envelope co-varied with the area of mouth opening, and Chandrasekaran *et al.*[2] who reported strong empirical evidence for Summerfield's claim. This suggests that intensity dynamics are strongly associated with articulatory movements that have direct influence on the speed by which the mouth opening area increases and decreases. Provided that this relationship exists and given the fact that articulatory movements vary between speakers, we hypothesize that intensity dynamics should also vary between speakers. This hypothesis was tested in the present experiment.

Why should we separate the intensity contour into positive and negative intensity dynamics? Birkholz *et al.*[14] examined the coordination between articulators by

[a]Author to whom correspondence should be addressed.

modelling both opening and closing gestures using dynamic systems. Opening and closing gestures are the articulatory movements to and from an articulatory target (typically a major turning point of articulators within a syllable). They found that the forces and motor programs acting on them in opening and closing gestures differed by their time constants. According to Ghez and Krakauer's[15] view of the motor program, the extent of a movement is planned before the movement is initiated. Speakers are therefore likely to have different motor planning for opening and closing gestures. Since the two gestures have different motor programs, it is unclear what effects this would have on the variability between speakers. For this reason, we looked at speaker-specific effects in positive and negative dynamics separately.

A series of measures was developed to capture how positive and negative intensity dynamics were distributed within utterances (Sec. 2.3). With them, we first tested whether measures of both dynamics formed into independent categories. Then, we tested whether and to what extent measures of both dynamics varied between speakers.

Why do we want to better understand speaker idiosyncratic temporal properties of the intensity contour? On the one hand there is a large theoretical interest. While indexical information has been deemed a by-product in classic linguistic theory, it is now evident that it plays a crucial role for the processing of meaning in speech communication.[16] The processes by which listeners recognize or distinguish different voices, however, are still poorly understood. Intensity contours might be factors contributing to auditory speaker recognition that have so far received hardly any attention. On the other hand, there are a variety of applications in which indexical information is of importance. In forensic voice analysis, for example, speaker comparison tasks often cannot be performed because the complexity of the acoustic correlates of voice identity within and between speakers is not yet well understood. It is thus essential to increase our knowledge beyond the classic factors like fundamental and formant frequencies or voice qualities to other acoustic domains that carry speaker-specific variation.

2. Method

2.1 Corpus

The TEVOID corpus[5,6] was used for the present study. It contains 16 native speakers of Zürich German (8 female, 8 male; mean age = 27, age standard deviation = 3.6, age range = 20–33, no reported speech and hearing disorders). They were recorded reading the same set of 256 sentences (see Fig. 1 for the distribution of sentence lengths in terms of syllable numbers) in a sound-attenuated booth (Neumann STH–100 transducer microphone (Georg Neumann GmbH, Berlin, Germany); 44.1k samples/s, 16-bit). All speakers practiced the sentences in advance to be able to read them fluently. The speakers read the sentences in a way they considered "everyday reading." Mm. 1 and Mm. 2 contain the sound files of the same sentence read by a female and a male speaker. Syllable boundaries were annotated automatically based on segment sonority rules; sonority scales were manually attributed to each segment type.[5,6]

Mm. 1. A female speaker reading the Zürich German sentence "Ich bin wäge Spraachwüsenschaft dänn usegheit." This is a file of type "wav" (266 Kb).

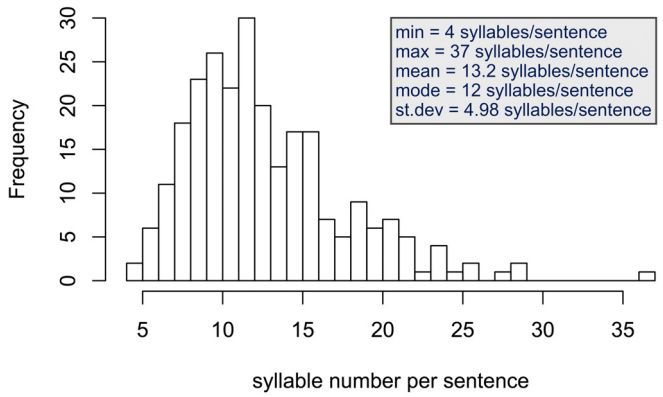

Fig. 1. (Color online) Histogram showing the distribution of sentence lengths (number of syllables per sentence) of the sentences in the TEVOID corpus.

Mm. 2. A male speaker reading the Zürich German sentence "Ich bin wäge Spraachwüsenschaft dänn usegheit." This is a file of type "wav" (291 Kb).

2.2 Extraction of the intensity contour and its peaks and troughs

We extracted the intensity contours to calculate the intensity dynamics measures (Sec. 2.3). First, the DC bias of each signal was removed by subtracting the mean amplitude. Then, the amplitude of each signal was linearly rescaled such that the maximum amplitude equated to 0.99. To obtain the intensity contour, the amplitude values of the rescaled signal were squared. A Gaussian window (approximated using the Kaiser-Bessel window: $\beta = 20$, sidelobe attenuation $\cong -190$ dB) with a length of 32 ms was multiplied repeatedly with the squared signal (window forward $= \frac{1}{4} \times 32\,\text{ms} = 8\,\text{ms}$; between-window overlap $= 75\%$). For each windowed frame, the sum of squares (SS) of the sample values was computed and substituted in $10\log_{10}\{[SS / (2 \times 10^{-5})]^2/0.032\}$ to obtain the intensity level (unit: dB re 20 μPa) in each particular frame.

Since the intensity curve obtained this way was a lower sampled function, we calculated the peak and trough points from the higher sampled amplitude envelope (obtained by low-pass filtering the full-wave rectified signal at 10 Hz [Hann filter, roll-off = 6 dB/octave]). Peak points (t_P in Fig. 2) were positioned where the envelope reached maximum values between syllable boundaries. Trough points (t_T in Fig. 2) were placed where the envelope reached minimum values between adjacent peak points. The intensity values at each peak and trough points (I_P and I_T in Fig. 2) were obtained from the intensity curve at each t_P and t_T using the cubic interpolation.

2.3 Measurement of intensity dynamics

Peak and trough points (t_P and t_T) and their associated intensity values (I_P and I_T) were obtained from each utterance. Positive dynamics ($v_I[+]$) were defined as $v_I[+] \stackrel{\text{def}}{=} (I_P - I_T)/(t_P - t_T)$, where I_P and I_T refer to the intensity values at peak and trough points represented by t_P and t_T. Similarly, negative dynamics ($v_I[-]$) were defined as $v_I[-] \stackrel{\text{def}}{=} |I_T - I_P|/(t_T - t_P)$. Absolute values were taken because we were only interested in the magnitude. Thus, we measured the speed of intensity increases and decreases. Geometrically, $v_I[+]$ and $v_I[-]$ can be demonstrated as the secant lines $\overrightarrow{I_T I_P}$ and $\overrightarrow{I_P I_T}$ in Fig. 2, and we measured the steepness of these lines.

To capture the distributions of both types of dynamics in an utterance, mean, standard deviation, and Pairwise Variability Index (PVI; for a tuple Q with n elements $\{q_1, q_2, \cdots, q_n\}$, the PVI of $Q = \sum_{i=1}^{n-1} |q_i - q_{i+1}|/(n-1)$) of both positive and negative dynamics were calculated. The PVI calculates the averaged differences between consecutive acoustic magnitudes in a speech signal (e.g. temporal intervals or here intensity dynamics).[17] It was demonstrated to be particularly suitable for summarizing the sequential variability in speech over the course of an entire utterance.[5–7,17] We notated these measures as MEAN_ $v_I[+]$, STDEV_ $v_I[+]$ and PVI_ $v_I[+]$ for positive dynamics, and MEAN_ $v_I[-]$, STDEV_ $v_I[-]$ and PVI_ $v_I[-]$ for negative dynamics. They represented

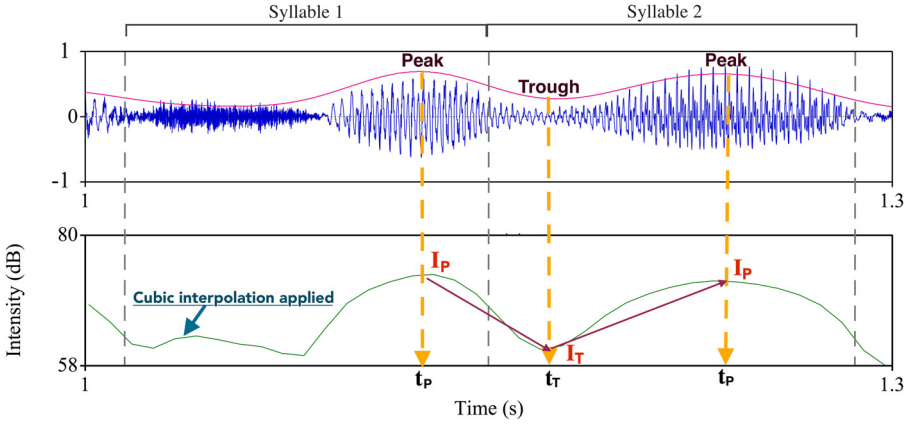

Fig. 2. (Color online) An illustration of calculating positive and negative intensity dynamics from a speech signal. The intensity contour (lower plot) was calculated from the speech waveform (upper plot). The amplitude envelope (superimposed over the waveform in the upper plot) was used to facilitate locating the peak and trough points (t_P and t_T). The peak and trough intensity values (I_P and I_T) were obtained from the intensity contour at t_P and t_T using the cubic interpolation. Intensity dynamics were calculated as how fast the intensity level dropped from a peak to its adjacent trough ($\overrightarrow{I_P I_T}$ in the lower plot, i.e., negative dynamics), or increased from a trough to its adjacent peak ($\overrightarrow{I_T I_P}$ in the lower plot, i.e., positive dynamics).

different aspects of dynamic distributions: i.e., the central tendency, the overall dispersion and sequential variability.

2.4 Statistical analyses

To control for the effect of between-sentence differences, z-score normalizations by sentence were performed for all measures of intensity dynamics: for a particular measure, the z-score of a particular sentence k was calculated as $z_k = (y_k - \bar{y}_k)/\sigma_k$, where y_k = the raw score of sentence k, $\bar{y}_k$ = the mean, and σ_k = the standard deviation of all y_k.

To test whether measures of positive and negative dynamics formed into independent categories, we performed a factor analysis (extraction method = principal components, eigenvalues ≧ 1, rotation method = Varimax with Kaiser normalization) on all measures of intensity dynamics. If measures in the two types of dynamics were classified as separate factors, we concluded that they were orthogonal and therefore encode different information.

To test the significance of between-speaker effect on each measure of intensity dynamics and the amount of between-speaker variation explained by measures of both dynamics, we employed a multinomial logistic regression (MLR). Measures of intensity dynamics were modeled as the numeric predictor variables, and speaker was modeled as the nominal response variable. Between-speaker variability explained by each measure was calculated as $(\chi^2/\Sigma\chi^2)\times 100\%$, where χ^2 refers to the likelihood ratio χ^2 of a particular measure, and $\Sigma\chi^2$ refers to the sum of likelihood ratio χ^2s of all measures.

3. Results

3.1 Factor analysis

The Kaiser-Meyer-Olkin (KMO) measure of sampling adequacy ($KMO = 0.669 > 0.5$) and the Bartlett's sphericity test ($\chi^2_{[15]} = 13249.911$, $p < 0.0005$) indicated that our dataset was suitable for factor analysis. Table 1 shows that two factors were extracted: factor 1 included all measures of negative dynamics and factor 2 included all measures of positive dynamics, suggesting that measures of both dynamics types were orthogonal.

3.2 Multinomial logistic regression

Table 2 shows the results of the MLR, examining the significance of between-speaker effect on each measure of intensity dynamics and how much between-speaker variability was explained by each measure. The negative measures collectively explained 70.35%, and the positive measures collectively explained 29.65% of between-speaker variability [Fig. 3(a)]. Figure 3(b) compares the difference between dynamics within each type of measure in explaining between-speaker variability.

4. Discussion

This paper investigated macroscopic intensity dynamics in the speech signal. Results from the MLR largely conformed to the hypothesis that intensity dynamics vary between speakers by showing that the between-speaker effect was significant in almost all measures of intensity dynamics, except PVI_ v_I[+] (see Table 2). Additionally, the amount of between-speaker variability explained by measures of both dynamics was not balanced: around 70% of between-speaker variation was explained by measures of negative dynamics. What could such a result tell us? Positive and negative dynamics

Table 1. Factor loadings matrix after Varimax rotation. The shaded loading values indicate that they are greater than the threshold (0.40), hence their associated intensity dynamics measures are classified into a particular factor.

	Factor loadings[a]	
	Factor 1	Factor 2
MEAN_ v_I[–]	0.825	0.043
STDEV_ v_I[–]	0.929	0.025
PVI_ v_I[–]	0.904	0.033
MEAN_ v_I[+]	0.098	0.780
STDEV_ v_I[+]	–0.008	0.926
PVI_ v_I[+]	0.003	0.908
Eigenvalue	2.497	2.169
% of variance explained	41.613	36.157

[a]The absolute value of a loading smaller than 0.40 indicates that the particular measure has an ignorable contribution to explaining the variance of a particular factor, and should therefore not be classified into this factor.

Table 2. Results of multinomial logistic regression.

	−2LL	$\chi^2_{[df]}$[a]	p	Variability explained[b]
(i) Model fitting information				
Null model	22713.047			
Full model	19958.848	2754.199 $_{[90]}$	<0.0005	
(ii) Likelihood ratio test of each measure of intensity dynamics				
MEAN_ v_I[−]	20907.008	948.161 $_{[15]}$	<0.0005	59.38%
STDEV_ v_I[−]	20100.527	141.679 $_{[15]}$	<0.0005	8.88%
PVI_ v_I[−]	19992.198	33.351 $_{[15]}$	<0.004	2.09%
MEAN_ v_I[+]	20304.375	345.527 $_{[15]}$	<0.0005	21.64%
STDEV_ v_I[+]	20064.253	105.406 $_{[15]}$	<0.0005	6.60%
PVI_ v_I[+]	19981.516	22.668 $_{[15]}$	= 0.09	1.42%
		$\Sigma\chi^2$ =1596.792		Σ% = 100%

[a]The χ^2 value of the final model was calculated by taking the difference between the −2log-likelihood ratios (–2LL) of the null model and the final model. The χ^2 value of each tested measure was calculated by taking the difference between the −2LLs of the final model and each reduced model.

[b]The variability explained was calculated by taking the percentage of the χ^2 value of each measure over the sum of all χ^2 values for all measures ($\Sigma\chi^2$).

might to some degree be influenced by opening and closing gestures, respectively, and thus carry two different types of information: the opening gestures might be more prosodically controlled as they may contain more information that is functional in linguistic terms, while the closing gestures might contain more speaker-specific information. According to the motor program theory, the central nervous system of the speaker actively plans and controls the articulatory behaviors in order to reach articulatory targets.[14,15] It seems plausible that such targets co-occur with mouth opening turning points which again co-occur with vocalic intensity peaks in the acoustic signal. To maximize mutual intelligibility, speakers of the same language should behave more similarly while reaching the same target. Once the target has been reached, the speaker may reduce the degree of control over the articulators, thereby producing movements which are determined more by the ontogenetic biophysical properties (e.g., the mass, damping, and friction) of their bones and muscles. In other words, these two processes are possibly influenced by two properties of the motor plant: controllable properties and intrinsic properties.[11] We argue that the controllable properties play a larger role in the opening gestures, while the intrinsic properties play a larger role in the closing gestures.

Our findings may be of particular interest to research where the identity information about a speaker matters, such as forensic phonetics and automatic speaker recognition. Our results showed that negative dynamics reveal more between-speaker variability than positive dynamics. This means that different parts of the signal intensity contour are more suitable for obtaining speaker-specific information. As such, these parts of the contour might be particularly relevant for forensic speaker comparisons or automatic speaker recognition. A related approach has been shown by Adami *et al.*[12] who fitted a single regression line over the entire energy contour of each syllable to model speaker individuality. The model may perform even better if features pertaining to negative dynamics were included. The theoretical implications of our findings, in particular the assumed relationship between articulatory movements and intensity dynamics requires further in-depth research:

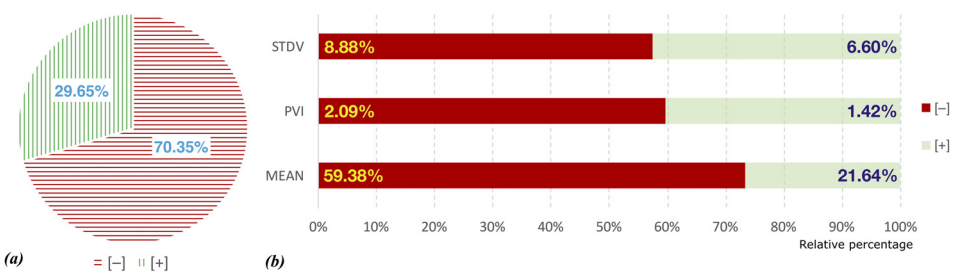

Fig. 3. (Color online) (a) Pie chart showing the amount of between-speaker variability explained by measures of positive dynamics ([+], vertical lines) and negative dynamics ([−], horizontal lines), respectively. (b) Stacked bar chart illustrating relative contributions of both dynamics within the same types of measures; absolute contributions are shown in numbers in each bar.

- We need to take into consideration that there are a variety of factors contributing to the variability of intensity levels in speech. Apart from the size of mouth aperture, there are factors like vocal effort, inherent vowel intensity, prosodic stress and accent or phonotactic arrangements of consonant-vowel sequences. It is imperative that we learn more about the complex relationships between these factors and the actual role that individual movements leading to mouth aperture size play in the individuality of intensity contour characteristics.
- It will be essential to examine the relationships between intensity dynamics and articulatory behavior with articulatory measurement procedures in which the effects of the trajectories of a variety of articulators on the intensity contours are tested. It will also be crucial to learn from such articulatory measurements to what degree the possible articulatory movements contributing to intensity contour variability are intrinsic and to what degree they are acquired behaviors.
- To generalize our findings, replications of results with languages other than our test language (Zürich German) is necessary. Such languages should ideally have different phonological complexities like vowel reductions, consonantal cluster complexities or word stress or accent variability that all might have an impact on articulatory movements and intensity contours.
- So far, we have studied rehearsed read speech only. It seems plausible that articulatory movements are more tensely controlled when the speech needs to be planned during the production process like in spontaneous speech. This speech is also characterized by hesitations, false starts and filled pauses which might have a strong influence on articulatory control.[18]
- Further research is needed to examine, for example, how intensity contours are affected by different forms of signal distortions, especially distortion that can directly affect amplitude envelopes non-linearly, such as dynamic range compressions.

Acknowledgments

This research was supported by the Gebert Rüf Stiftung (No. GRS-027/13) and the Swiss National Science Foundation (No. 100015_135287). We wish to thank Richard Rhodes for helpful suggestions on earlier drafts of the paper, and Adrian Leemann and Marie-José Kolly for their work in constructing the corpus.

References and links

[1]V. Dellwo, M. Huckvale, and M. Ashby, "How is individuality expressed in voice? An introduction to speech production and description for speaker classification," in *Speaker Classification I: Fundamentals, Features and Methods*, edited by C. Müller (Springer, Berlin, Germany, 2007), pp. 1–20.

[2]C. Chandrasekaran, A. Trubanova, S. Stillittano, A. Caplier, and A. A. Ghazanfar, "The natural statistics of audiovisual speech," PLoS Comput. Biol. **5**, e1000436 (2009).

[3]A. Eriksson, "Aural/acoustic vs. automatic methods in forensic phonetic case work," in *Forensic Speaker Recognition: Law Enforcement and Counter-Terrorism*, edited by A. Neustein and H. A. Patil (Springer, New York, 2012), pp. 41–69.

[4]T. Kinnunen and H. Li, "An overview of text-independent speaker recognition: From features to supervectors," Speech Commun. **52**, 12–40 (2010).

[5]A. Leemann, M.-J. Kolly, and V. Dellwo, "Speaker-individuality in suprasegmental temporal features: Implications for forensic voice comparison," Forensic Sci. Int. **238**, 59–67 (2014).

[6]V. Dellwo, A. Leemann, and M.-J. Kolly, "Rhythmic variability between speakers: Articulatory, prosodic, and linguistic factors," J. Acoust. Soc. Am. **137**, 1513–1528 (2015).

[7]L. He and V. Dellwo, "The role of syllable intensity in between-speaker rhythmic variability," Int. J. Speech Language Law **23**, 243–273 (2016).

[8]R. Kanai and G. Rees, "The structural basis of inter-individual differences in human behavior and cognition," Nat. Rev. Neurosci. **12**, 231–242 (2011).

[9]D. A. Winter, *Biomechanics* and *Motor Control of Human* Movement, 4th ed. (John Wiley and Sons, Hoboken, NJ, 2009), 320 pp.

[10]P. Perrier and R. Winkler, "Biomechanics of the orofacial motor system: Influence of speaker-specific characteristics on speech production," in *Individual Differences in Speech Production and Perception*, edited by S. Fuchs, D. Pape, C. Petrone, and P. Perrier (Peter Lang, Frankfurt, Germany, 2015), pp. 223–254.

[11]P. Perrier, "Gesture planning integrating knowledge of the motor plant's dynamics: A literature review for motor control and speech motor control," in *Speech Planning and Dynamics*, edited by S. Fuchs, M. Weirich, D. Pape, and P. Perrier (Peter Lang, Frankfurt, Germany, 2012), pp. 191–238.

[12]A. Adami, R. Mihaescu, D. Reynolds, and J. J. Godfrey, "Modeling prosodic dynamics for speaker recognition," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing* (Hong Kong, 2003), pp. 788–791.

[13]Q. Summerfield, "Lipreading and audio-visual speech perception," Philos. Trans. R. Soc. London B **335**, 71–78 (1992).

[14]P. Birkholz, B. J. Kröger, and C. Neuschaefer-Rube, "Model-based reproduction of articulatory trajectories for consonant-vowel sequences," IEEE Trans. Audio Speech Language Processing **19**, 1422–1433 (2011).
[15]C. Ghez and J. Krakauer, "The organization of movement," in *Principles of Neural Science*, 4th ed., edited by E. R. Kandel, J. H. Schwartz, and T. M. Jessell (McGraw-Hill, New York, 2000), pp. 654–673.
[16]C. C. Creel and M. A. Tumlin, "On-line acoustic and semantic interpretation of talker information," J. Mem. Language **65**, 264–285 (2011).
[17]E. Grabe and E. L. Low, "Durational variability in speech and rhythm class hypothesis," in *Laboratory Phonology*, edited by C. Gussenhoven and N. Warner (Mouton de Gruyter, Berlin, Germany, 2002), Vol. 7, pp. 514–546.
[18]G. P. M. Laan, "The contribution of intonation, segmental durations and spectral features to the perception of a spontaneous and a read speaking style," Speech Commun. **22**, 43–65 (1997).